CROP ACREAGE ESTIMATION USING SATELLITE DATA AS AUXILIARY INFORMATION: MULTIVARIATE CASE

James J. McKeon and Raj S. Chhikara
University of Houston-Clear Lake

1. INTRODUCTION

The National Agricultural Statistical Service (NASS) of the U.S. Department of Agriculture (USDA) utilizes LANDSAT data to improve upon the regular USDA crop acreage estimates for several states. The multispectral scanner data acquired from the satellite are processed to obtain direct estimates of various crop acreages. These, in turn, are used as auxiliary variables and the regression estimates of crop acreages are obtained based on separate regressions. Because of crop competition, etc., the actual crop acreages for different crops in a stratum as well as their LANDSAT estimates are expected to be correlated. So it seems appropriate to consider the multivariate regression approach to crop acreage estimation when several crops are cultivated in a state.

In an earlier paper, Chhikara and Houston (1984) argued that in this application the auxiliary variable, in fact, is a dependant variable and the actual crop acreage is an independent variable in terms of the usual regression model set up. Though the regression model in this situation is linear, its order is reversed compared to a prediction model. McKeon, et.al. (1985) addressed this problem and investigated the regression estimator as well as two other estimators of a finite population mean. This investigation was for the univariate case only, involving a separate regression for each crop of interest. In this paper we extend it to the multivariate case and discuss the three multivariate estimators corresponding to those given previously in McKeon, et.al (1985) and Chhikara and McKeon (1985). Fuller (1986) has considered small area estimation including an extension to the multivariate case.

In general, the estimation problem may be stated as follows: Suppose the auxiliary vector ($\underline{x}$) is related to the response vector of interest ($\underline{y}$) by a linear model,

$$\underline{x} = \underline{\alpha} + \beta\underline{y} + \underline{\varepsilon} \qquad (1)$$

where

$$E[\underline{\varepsilon} \mid \underline{y}] = \underline{0}$$

and

$$E[\underline{\varepsilon}\,\underline{\varepsilon}' \mid \underline{y}] = \Sigma . \qquad (2)$$

There is a finite population of $\underline{y}$ values of size N, satisfying (1) with errors $\underline{\varepsilon}$ from an infinite population. Let $\bar{\underline{X}}$ and $\bar{\underline{Y}}$ be the mean vectors of the $\underline{x}$ and $\underline{y}$ populations. The problem is to estimate $\bar{\underline{Y}}$ when $\bar{\underline{X}}$ is known and a random sample of n paired vector observations $(\underline{x}_i, \underline{y}_i)$, i = 1,2,...,n, are given. It will be assumed that $\underline{x}$ and $\underline{y}$ are each a pxl vector.

2. ESTIMATORS

For the random sample $(\underline{x}_i, \underline{y}_i)$, i=1,..,n, from a finite population of size N, let $\bar{x}, \bar{y}$, be the sample means and let S_{xy}, S_{yy} and S_{xx} denote the sample covariance matrices based on n-1 degrees of freedom.

The parameter matrix β can be estimated by

$$B = S_{xy}S_{yy}^{-1}. \qquad (3)$$

The maximum likelihood estimator of the population mean $\bar{\underline{Y}}$, known as the classical estimator, is

$$\hat{\bar{\underline{Y}}}_C = \bar{\underline{y}} + B^{-1}(\bar{\underline{X}} - \bar{\underline{x}}). \qquad (4)$$

This estimator minimizes the measurement (model) error in (1) but has bias of O(1/n) and infinite variance.

2.1 Unbiased Classical Estimator

To obtain the conditional expectation of B^{-1} given $\underline{y}$, let

$$B = \beta + D \text{ where } E(D) = 0.$$

Then by a Taylor expansion

$$B^{-1} = \beta^{-1} - \beta^{-1}D\beta^{-1} + \beta^{-1}D\beta^{-1}D\ldots \qquad (5)$$

From the covariances between elements of a matrix of linear regression coefficients as given in Anderson (1958, p.182),

$$E(DB^{-1}D) = \Sigma\beta^{-1}S_{yy}/(n-1) \qquad (6)$$

and

$$E(B^{-1}) = (I+\Gamma S_{yy}^{-1}/(n-1))\,\beta^{-1} \qquad (7)$$

with $\Gamma = \beta^{-1}\Sigma\beta^{-1}$

An estimator of β^{-1} unbiased up to $O(1/n)$ is given by

$$(B^{-1})_u = [I+B^{-1}\hat{\Sigma}B^{-1}{}'S_{yy}/(n-1)]^{-1}B^{-1}$$
$$= S_{yy}B'[BS_{yy}B'+\hat{\Sigma}/(n-1)]^{-1} \quad (8)$$

Replacing B^{-1} in (4) by $(B^{-1})_u$ gives the modified classical estimator,

$$\hat{\underline{\bar{Y}}}_u = \underline{\bar{y}} + (B^{-1})_u(\bar{\underline{X}} - \underline{\bar{x}}), \quad (9)$$

which is conditionally unbiased to $O(1/n)$. Making use of the expansion for B^{-1} in (5), the conditional covariance matrix to $O(1/n^2)$ is given by

$$\text{Cov}\,(\hat{\underline{\bar{Y}}}_u|\underline{y}) = [(1-f)/n]\ \Gamma\,[1+(\text{tr}\,\Gamma S_{yy}^{-1}$$
$$+ n(\underline{\bar{Y}}-\underline{\bar{y}})'S_{yy}^{-1}(\underline{\bar{Y}}-\underline{\bar{y}}))/(n-1)] \quad (10)$$

Taking the expectation with respect to random sampling, the unconditional covariance matrix of $\hat{\underline{\bar{Y}}}_u$ is

$$\text{Cov}(\hat{\underline{\bar{Y}}}_u) = \Gamma[(1-f)/n][1+(p+\text{tr}\Gamma\Sigma_{yy}^{-1})/(n-p-2)]$$
$$+ O(1/n^3). \quad (11)$$

In obtaining (11), we have made use of the fact that if S_{yy} is based on a sample from a multivariate normal distribution,

$$E(S_{yy}^{-1}) = \Sigma_{yy}^{-1}(n-1)/(n-p-2) \quad (12)$$

(Haff, 1982).

For non-normal distributions, (12) is the first order term.

This estimator $\hat{\underline{\bar{Y}}}_u$ is unconditionally unbiased with

$$E(\hat{\underline{\bar{Y}}}_u) = \underline{\bar{Y}} + O(1/n^2). \quad (13)$$

2.2 Regression Estimator

On the other hand one may consider the usual regression estimator based on the regression of $\underline{y}$ on $\underline{x}$. This estimator is given by

$$\hat{\underline{\bar{Y}}}_R = \underline{\bar{y}} + T(\bar{\underline{X}} - \underline{\bar{x}}) \quad (14)$$

where

$$T = S_{xy}\,S_{xx}^{-1}. \quad (15)$$

Under random sampling the regression estimator minimizes the overall estimation error whereas the classical estimators minimize the model error.

Under general conditions, that is, without assuming normality or linear regression, the unconditional bias of $\hat{\underline{\bar{Y}}}_R$ can be obtained by letting

$$T = (\Sigma_{yx}+U_{11})(\Sigma_{xx}+U_{20})^{-1} \quad (16)$$

where U_{11} and U_{20} are random deviation matrices. Retaining only the linear terms in the expansion,

$$T = T_0 - T_0\,U_{20}\,\Sigma_{xx}^{-1} + U_{11}\,\Sigma_{xx}^{-1} + \ldots \quad (17)$$

with $T_0 = \Sigma_{yx}\,\Sigma_{xx}^{-1}$, (18)

the bias of $\hat{\underline{\bar{Y}}}_R$ to $O(1/n)$ is given by

$$E[T(\bar{\underline{X}}-\underline{\bar{x}})] = -[(1-f)/n]E[\underline{w}(\underline{\bar{x}}-\bar{\underline{X}})'\Sigma_{xx}^{-1}(\underline{\bar{x}}-\bar{\underline{X}})] \quad (19)$$

where

$$w_i = (\underline{y}_i-\underline{\bar{y}}) - T(\underline{x}_i-\underline{\bar{x}})$$
$$= (\underline{y}_i-\hat{\underline{\bar{Y}}}_R) - T(\underline{x}_i-\bar{\underline{X}}) \quad (20)$$

This bias is a function of the third moments of the $(\underline{x},\underline{y})$ distribution and is small or zero for near normal distributions.

We remark here that for a linear regression of y on x, when conditioning on $\underline{x}$, the regression estimator is unbiased. However, conditioning on $\underline{y}$, the regression estimator has a bias of order one, that is

$$\text{Bias}(\hat{\underline{\bar{Y}}}_R|\underline{y}) = (I - T\beta)(\underline{\bar{y}} - \underline{\bar{Y}}) + \ldots$$

Determination of $\text{Cov}\,(\hat{\underline{\bar{Y}}}_R)$ under model (1) without additional assumptions is highly involved and its asymptotic expression is not very useful. A generalization of Cochran's (1977) univariate variance formula for the regression estimator can be obtained under the assumption that the regression of $\underline{y}$ on $\underline{x}$ is linear with homogeneous errors. The conditional covariance matrix of $\hat{\underline{\bar{Y}}}_R$ to $O(1/n^2)$ is then

$$\text{Cov}(\hat{\underline{\bar{Y}}}_R|\underline{x}) =$$

$$[(1-f)/n]\ [\Sigma_{yy} - \Sigma_{yx}\,\Sigma_{xx}^{-1}\,\Sigma_{xy}]$$

$$[1+n(\bar{\underline{x}} - \bar{\underline{X}})'\,S_{xx}^{-1}(\bar{\underline{x}} - \bar{\underline{X}})/(n-1)]. \quad (21)$$

Taking the expectation over the sampling variation of $\underline{x}$ gives the uncondition covariance matrix,

$$\mathrm{Cov}(\hat{\bar{\underline{Y}}}_R) = [(1-f)/n]\ [\Sigma_{yy}^{-1} - \Sigma_{yx}\ \Sigma_{xx}^{-1}\ \Sigma_{xy}]$$

$$[1+p/(n-p-2)]. \qquad (22)$$

3. CONFIDENCE REGIONS FOR $\bar{\underline{Y}}$

The sample analogues of equations (11) and (22) can be used to obtain confidence regions for $\bar{\underline{Y}}$. The covariance estimates are obtained by replacing the parametric quantities β, Γ and Σ_{yy} by their corresponding sample estimates,

$$\hat{\beta} = B$$

$$\hat{\beta}^{-1} = (B^{-1})_u$$

$$\hat{\Gamma} = (B^{-1})_u\ \hat{\Sigma}\ (B^{-1})_u.$$

The matrix

$$\Sigma_{y|x} = [\Sigma_{yy} - \Sigma_{yx}\ \Sigma_{xx}^{-1}\ \Sigma_{xy}]$$

is estimated by (23)

$$[S_{yy}-S_{yx}S_{xx}^{-1}S_{xy}]\ (n-1)/(n-p-1)$$

and $\Sigma = \Sigma_{x|y}$ is similarly estimated.

Letting $\hat{\bar{\underline{Y}}}$ with covariance matrix V_y represent either $\hat{\bar{\underline{Y}}}_u$ or $\hat{\bar{\underline{Y}}}_R$, with $\hat{\bar{\underline{Y}}}_R$ unbiased, an approximate $(1-\alpha)$ level confidence ellipsoid for $\bar{\underline{Y}}$ is given by

$$(\hat{\bar{\underline{Y}}} - \bar{\underline{Y}})'\ \hat{V}_y^{-1}\ (\hat{\bar{\underline{Y}}} - \bar{\underline{Y}}) \leq T^2_{n-p-1,\alpha} \qquad (24)$$

where $T^2_{n-p-1,\alpha}$ is the $(1-\alpha)$ percentage point of a p-variate Hoteling's T^2 with n-p-1 degrees of freedom for error. When there are more than two variables, the confidence region in (24) becomes difficult to interpret. Simultaneous confidence intervals for the components of $\bar{\underline{Y}}$ can be obtained by the Scheffe method or by applying the Bonferroni inequality which, in this application, will give shorter intervals.

4. CORRELATION STRUCTURE IN CROP AREA ESTIMATION

The pattern of the $(\underline{y},\underline{x})$ correlation matrix determines whether the multivariate approach provides any gain over the univariate. Introducing the simplifying assumption of symmetry between the $\underline{x}$ variables and between the $\underline{y}$ variables, let

ρ_a = correlation between paired variables (y_i,x_i), i = 1 to p,

ρ_b = correlation between unpaired variables (y_i, x_j), $i \neq j$,

ρ_x = correlation between $\underline{x}$ variables,

ρ_y = correlation between $\underline{y}$ variables

For separate univariate estimation, the variance of the regression estimator is

$$V_1(\hat{\bar{\underline{Y}}}_R) = [(1-f)/n](1 - \rho_a^2)(1 + \frac{1}{n-3}), \qquad (25)$$

and for multivariate estimation,

$$V_m(\hat{\bar{\underline{Y}}}_R) = [(1-f)/np]\ \ [1 + p/(n-p-2)]$$
$$[p - (p-1)(\rho_a-\rho_b)^2/(1-\rho_x)$$
$$- (\rho_a+(p-1)\ \rho_b)^2/(1+(p-1)\ \rho_x)]. \qquad (26)$$

By substitution into (26), if

$$\rho_b = \rho_a\ \rho_x \quad , \qquad (27)$$

the univariate estimation variance is always less than the multivariate variance, that is

$$V_1(\hat{\bar{\underline{Y}}}_R) < V_m(\hat{\bar{\underline{Y}}}_R), \text{ for all n.}$$

For a general correlation matrix the condition

$$\rho_{y_i x_j|x_i} = 0 \quad \text{if and only if}$$
$$\rho_{y_i x_j} = \rho_{y_i x_i}\ \rho_{x_i x_j},\ i,j = 1 \text{ to } p, \qquad (28)$$

results in $V_1(\hat{\bar{\underline{Y}}}_R) < V_m(\hat{\bar{\underline{Y}}}_R)$, for all n.

5. SIMULATIONS COMPARING UNIVARIATE AND MULTIVARIATE ESTIMATION

Four measures were used to compare the multivariate vs. univariate regression estimators:

1) Closeness - the number of times the estimator was closer to the true $\bar{\underline{Y}}$ vector than the alternative estimator.

2) Box - the number of times the estimator was within a fixed box centered at $\bar{\underline{Y}}$. The size of the box was choosen so that approximately 90% of the estimates would be within the box.

3) MSE - mean squared error of the estimator averaged over all replications.

4) Interval width - Applying the Bonferoni inequality, the 90% confidence ellipsoid for each estimate was converted to simultaneous confidence intervals. The interval widths were averaged over the p variates and over replications.

The ratio of multivariate to univariate values for the four measures are used to assess the relative efficiency of the estimators. For the two crop case (corn and soybeans), acreage estimates for 16 area segments taken from Harter (1983), were pooled to estimate the 4 X 4 covariance matrix of $(\underline{y}, \underline{x})$. Simulations based on these correlation values were carried out for sample size of 10, 15, 25 and 40 with 1000 replications. Values for "Closeness" and "Box" greater than one and values for "MSE" and "Width" less than one indicate an improvement using multivariate in place of univariate estimation. Table 1a. shows a slight gain using multivariate estimation for the two crop case. This may be the result of the small number (16) of segments upon which the correlations were based upon, allowing random deviations from the condition stated in (28).

A case using data for four crops was studied. The covariance matrix for the four crop case is based on simulated segment data. Using 20 data points, results are similar to the two crop case, showing a slight advantage for multivariate estimation. When the number of data points is increased to 80, thereby significantly reducing the random deviations from the condition stated in (28), the univariate estimation is uniformly better even for a sample size of 40 (Table 1b). The apparent advantage of multivariate estimation in the two crop case is mostly due to the small number of segments used in estimating the correlation matrix.

The multivariate approach will show an advantage only when information about y_i is contained in some non-paired variable x_j, $j \neq i$. Table 2. shows the results of a small (30^0) rotation of the pair (x_1, x_2) and the pair (x_3, x_4). Multivariate estimation now provides a substantial improvement on univariate, especially evident in the four crop case.

Table 1. Comparison of Estimation Methods

(a) Corn and Soybeans

(Actual data for 16 segments)

RMS multiple corr. = .79, RMS paired corr. = .75

Correlation matrix:

y_1	1.00	-.21	.70	-.37
y_2		1.00	-.31	.80
x_1			1.00	-.67
x_2				1.00

Ratios for multivariate vs. univariate estimation

NS	Closeness	Box	MSE	Width
10	1.22	1.00	.97	1.02
15	1.17	.99	.96	.96
25	1.20	1.00	.91	.95
40	1.38	1.02	.84	.92

(b) Four Crop Types

(80 Simulated data points)

RMS multiple corr. = .72, RMS paired corr. = .70

Correlation matrix: (omitted)

Ratios for multivariate vs. univariate estimation

NS	Closeness	Box	MSE	Width
10	.43	.87	1.69	1.46
15	.62	.92	1.30	1.16
25	.73	.95	1.12	1.06
40	.94	.99	1.03	1.02

Table 2. Comparisons with Correlation Matrices Modified by Rotating Adjacent Pairs

(a) Corn and Soybeans

RMS multiple corr. = .79, RMS paired corr. = .75

Correlation matrix:

y_1	1.00	-.21	.82	-.51
y_2		1.00	.14	.68
x_1			1.00	.15
x_2				1.00

Ratios for multivariate vs. univariate estimation

NS	Closeness	Box	MSE	Width
10	2.85	1.11	.57	.72
15	2.89	1.13	.56	.71
25	3.10	1.14	.53	.70
40	3.48	1.18	.46	.64

(b) Four Crop Types

RMS Multiple Corr. = .72, RMS paired corr. = .64

Correlation matrix: (omitted)

Ratios for multivariate vs. univariate estimation

NS	Closeness	Box	MSE	Width
10	1.93	.97	.78	.91
15	3.26	1.03	.54	.69
25	3.88	1.04	.51	.66
40	4.71	1.06	.49	.65

ACKNOWLEDGEMENT

This research is supported under a grant from the National Agricultural Statistical Service of the U.S. Department of Agriculture.

REFERENCES

Chhikara, R.S. and Houston, A.G. (1984). "Calibration of inverse regression: which is appropriate for crop surveys using Landsat data?" Proceedings fo the Second Annual NASA Symposium on Mathematical Pattern Recognition and Image Analysis, Johnson Space Center, Houston, p. 205-243.

Chhikara, R.S. and McKeon, J.J. (1985), "Estimation of finite populations under a calibration model," Unpublished Manuscript, Univ. of Houston-Clear Lake, Houston, Texas

Cochran, W.G. (1977). Sampling Techniques. John Wiley & Sons, New York.

Fuller, W.S. (1986). "Small area estimation as a measurement error problem," Unpublished Manuscript, Department of Statistics, Iowa State University, Ames, Iowa.

Haff, L.R. (1982). "Identifies for the inverse Wishart distribution," Sankhya, Series B, 44, p. 245-258.

Harter, Rachel M. (1983). "Small area estimation using rested-error models and auxiliary data", Ph.D. Thesis, Iowa State University, Ames, Iowa.

McKeon, J.J., Chhikara, R.S. and Boullion,T. (1985), ""Linear regression estimators in sample surveys under callibration." ASA 1985 Proceedings of the Section on Survey Research Methods, p. 286-290.

McKeon, J.J. (1985), "Statistical Calibration Theory," Ph.D. Thesis, Old Dominion University, Norfolk, Virginia.